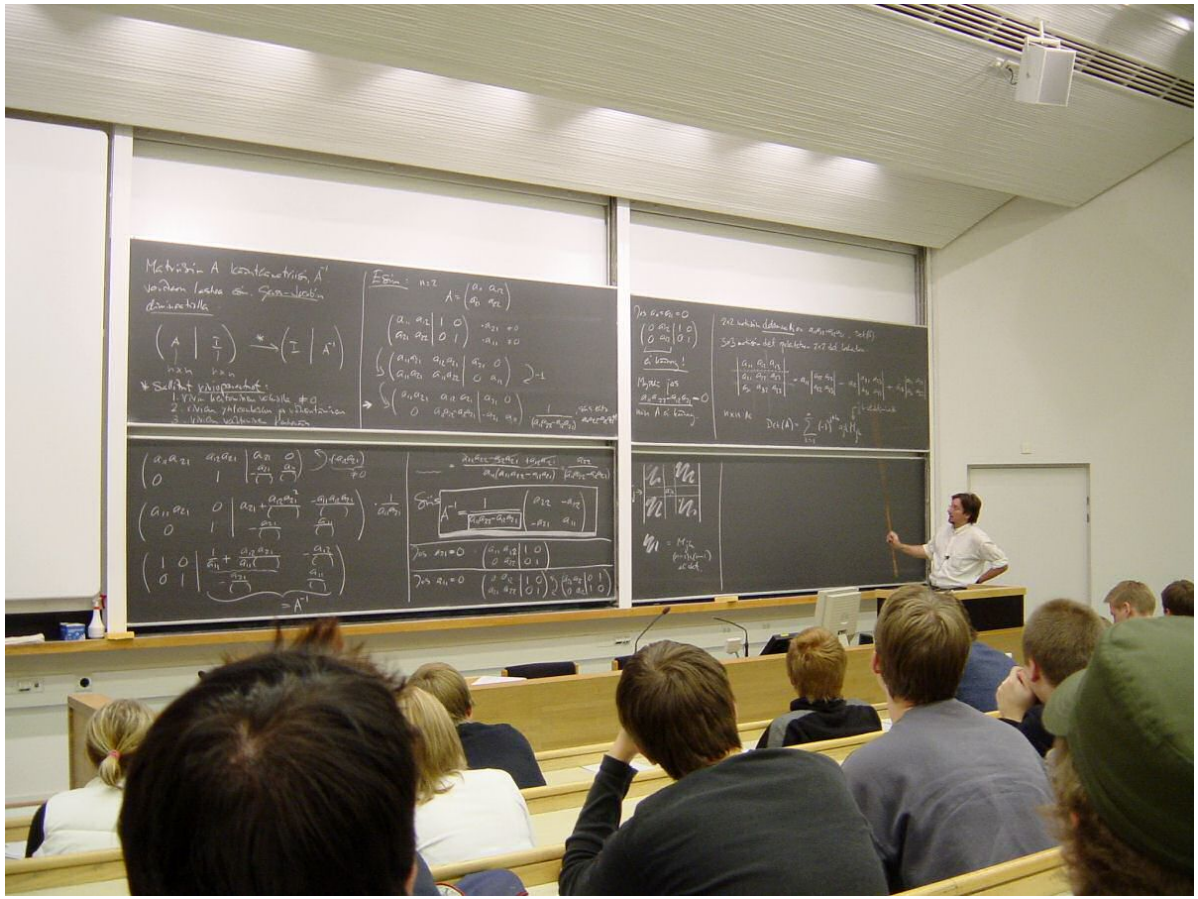
What is...machine learning in mathematics - part 13?

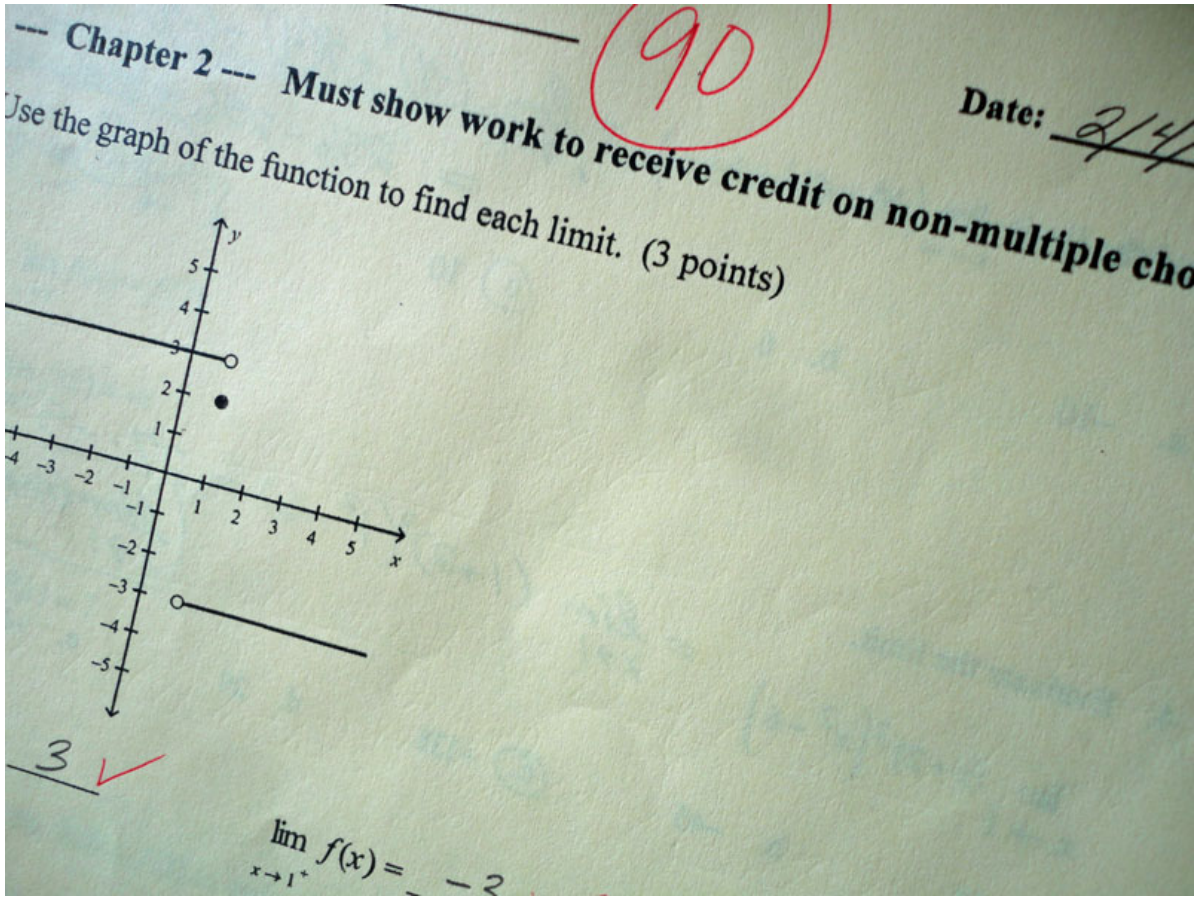
Or: Grading math with AI!?

Mathematics education



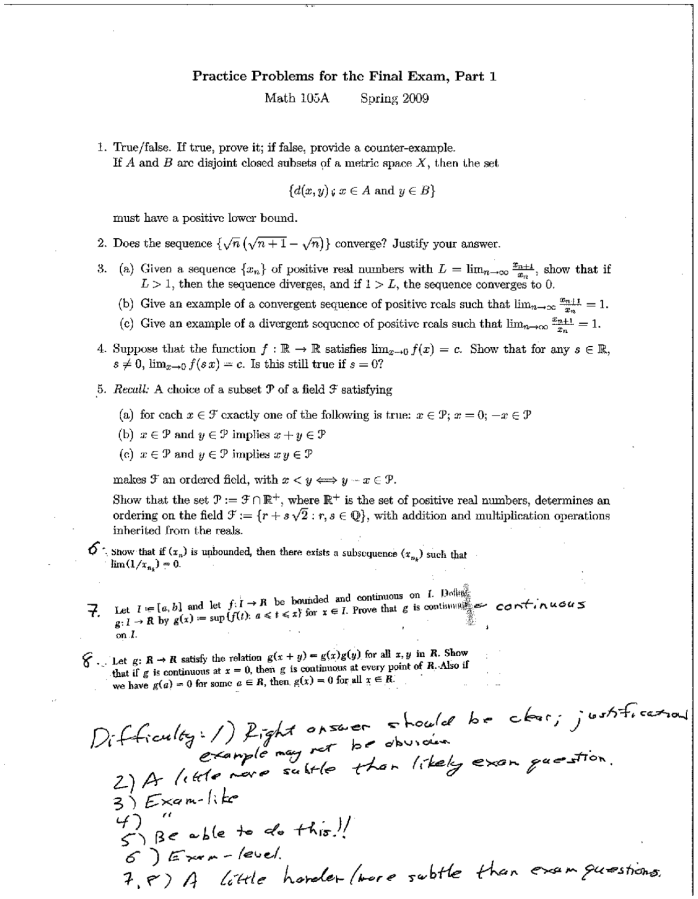
- Above $\approx$ how I learned linear algebra
- Question How can machine learning (ML) help in math education?
- This video us mostly about university level

Grading



- The question is way too big so let us zoom into grading
- Grading = a necessary evil that nobody (?) likes
- Grading is a prototypical example of what ML should replace

Easier (?) than research



Practice Problems for the Final Exam, Part 1

Math 105A Spring 2009

1. True/false. If true, prove it; if false, provide a counter-example.
If A and B are disjoint closed subsets of a metric space X, then the set

$$\{d(x,y)\,;\, x \in A \text{ and } y \in B\}$$

must have a positive lower bound.

2. Does the sequence $\{\sqrt{n}\,(\sqrt{n+1}-\sqrt{n})\}$ converge? Justify your answer.

3. (a) Given a sequence $\{x_n\}$ of positive real numbers with $L = \lim_{n\to\infty} \frac{x_{n+1}}{x_n}$, show that if $L > 1$, then the sequence diverges, and if $1 > L$, the sequence converges to 0.
(b) Give an example of a convergent sequence of positive reals such that $\lim_{n\to\infty} \frac{x_{n+1}}{x_n} = 1$.
(c) Give an example of a divergent sequence of positive reals such that $\lim_{n\to\infty} \frac{x_{n+1}}{x_n} = 1$.

4. Suppose that the function $f : \mathbb{R} \to \mathbb{R}$ satisfies $\lim_{x\to 0} f(x) = c$. Show that for any $s \in \mathbb{R}$, $s \neq 0$, $\lim_{x\to 0} f(sx) = c$. Is this still true if $s = 0$?

5. *Recall:* A choice of a subset $\mathcal{P}$ of a field $\mathcal{F}$ satisfying
(a) for each $x \in \mathcal{F}$ exactly one of the following is true: $x \in \mathcal{P}$; $x = 0$; $-x \in \mathcal{P}$
(b) $x \in \mathcal{P}$ and $y \in \mathcal{P}$ implies $x + y \in \mathcal{P}$
(c) $x \in \mathcal{P}$ and $y \in \mathcal{P}$ implies $xy \in \mathcal{P}$
makes $\mathcal{F}$ an ordered field, with $x < y \iff y - x \in \mathcal{P}$.
Show that the set $\mathcal{P} := \mathcal{F} \cap \mathbb{R}^+$, where $\mathbb{R}^+$ is the set of positive real numbers, determines an ordering on the field $\mathcal{F} := \{r + s\sqrt{2} : r, s \in \mathbb{Q}\}$, with addition and multiplication operations inherited from the reals.

6. Show that if (x_n) is unbounded, then there exists a subsequence (x_{n_k}) such that $\lim(1/x_{n_k}) = 0$.

7. Let $I = [a,b]$ and let $f : I \to R$ be bounded and continuous on I. Define $g : I \to R$ by $g(x) = \sup\{f(t) : a \leq t \leq x\}$ for $x \in I$. Prove that g is continuous on I.

8. Let $g : R \to R$ satisfy the relation $g(x+y) = g(x)g(y)$ for all x, y in R. Show that if g is continuous at $x = 0$, then g is continuous at every point of R. Also if we have $g(a) = 0$ for some $a \in R$, then $g(x) = 0$ for all $x \in R$.

Difficulty: 1) Right answer should be clear; justification example may not be obvious
2) A little more subtle than likely exam question.
3) Exam-like
4) "
5) Be able to do this!!
6) Exam-level.
7, 8) A little harder/more subtle than exam questions.

- Grading math is similar to math research: arguments and answers matter
- However, in contrast to research, the answers are known
- A transformer should perform quite well, especially when working in collaboration with humans

Enter, the theorem

A (transformer) neural network (NN) did the following:

We use OpenAI's recent GPT-4o model to assign scores to student responses. We prompt the model to grade one question at a time, providing a scanned image of the corresponding page from the student's exam and telling the model how many points each part is worth. We experiment with 3 different prompt types: i) no context **(N)**, where the model only sees the student response, ii) correct answer **(C)**, where the model sees the student response and the correct answer for each question part, and iii) correct answer and rubric **(CR)**, where the model sees the student response, the correct answer, and the rubric for each question part. We measure how well GPT-4o can score student responses, which we refer to as alignment, by comparing its predicted scores to the ground truth scores assigned by course graders. We examine scores at the question level, resulting in $18 \times 5 = 90$ samples, and normalize scores between 0 and 1 based on the total points per question. We then compute the mean absolute error **(MAE)**, root mean squared error **(RMSE)**, accuracy **(Acc.)**, and Pearson's correlation coefficient **(Corr.)** between predicted and ground truth question scores. We also show the average score assigned by graders **(Score G.)** and by the model **(Score M.)**.

3 Results

Table 1: Average alignment by prompt type. Providing the answer and rubric performs the best.

Prompt Type	MAE ↓	RMSE ↓	Acc. ↑	Corr. ↑	Score G.	Score M.
N	0.0940	0.1533	0.4222	0.2776	0.8988	0.9759
C	0.0989	0.1609	0.4333	0.5502	0.8988	0.8501
CR	0.0766	0.1267	0.4667	0.6174	0.8988	0.8808

Major issues: reading the student's handwriting; GPT-4o is too diplomatic

- Note that the used NN is not fine tuned for grading
- My interpretation In 2025, automated grading proves most effective when used in conjunction with human input, much like for research questions

My biased conclusion

- In 2025 ML and NN in math can be used efficiently when combined with human expertise
- Works great Formal proof verification, pattern recognition, generating counterexamples, ...
- Reasoning seems to be still missing

Thank you for your attention!

I hope that was of some help.